Tufts
UNIVERSITY

SCHOOL OF ENGINEERING
Computer Science

Using Positive, Unlabeled, and Hard Negative Examples for Hierarchical Classification of Enzyme Promiscuity

Gian Marco Visani

Laidlaw Advisor: Soha Hassoun

Collaborator: Michael Hughes

Binary Classification

Cat

Dog

Training

Machine Learning predictor

Prediction

Machine Learning predictor

Cat

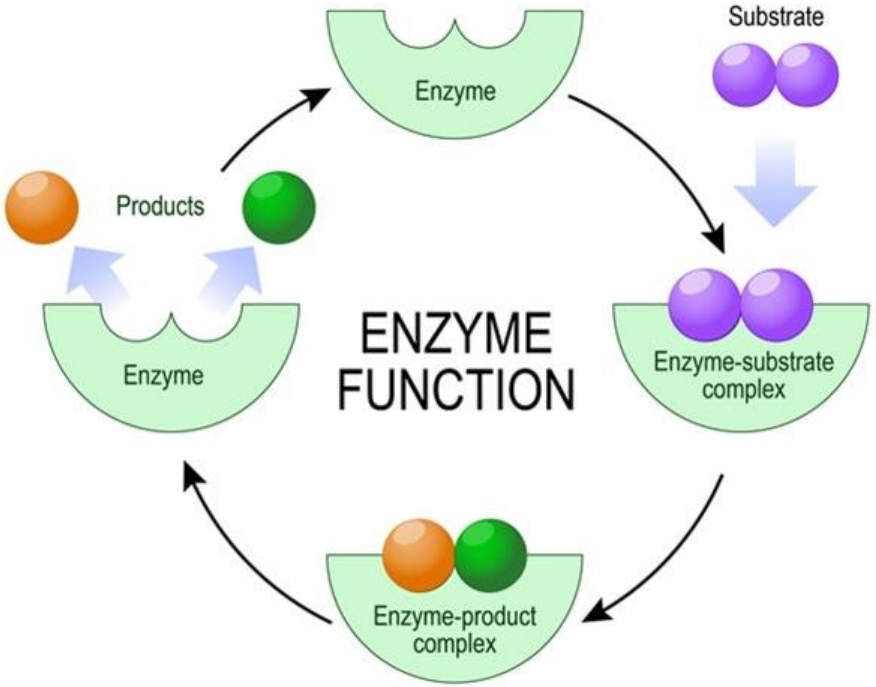
Courtesy: https://stockphoto.com

Apply classification

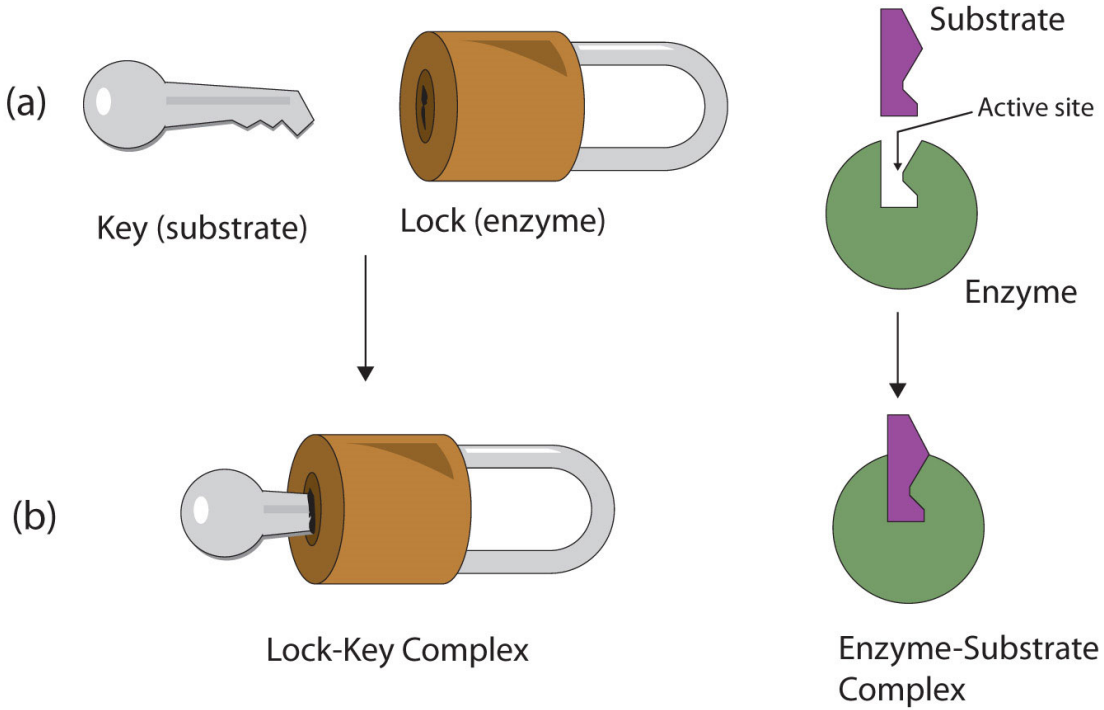
to predict enzyme activity

on any given molecule

How enzymes work



https://www.news-medical.net/life-sciences/What-is-an-Enzyme-Cofactor.aspx



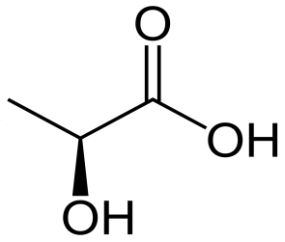
https://saylordotorg.github.io/text_the-basics-of-general-organic-and-biological-chemistry/s21-06-enzyme-action.html

Substrate Promiscuity

L-lactate dehydrogenase (EC 1.1.1.27)

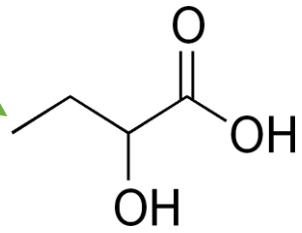
Natural

Lactic acid

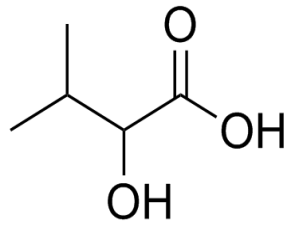


Promiscuous

2-hydroxybutyrate



2-hydroxy-3-methylbutanoate



Many more…

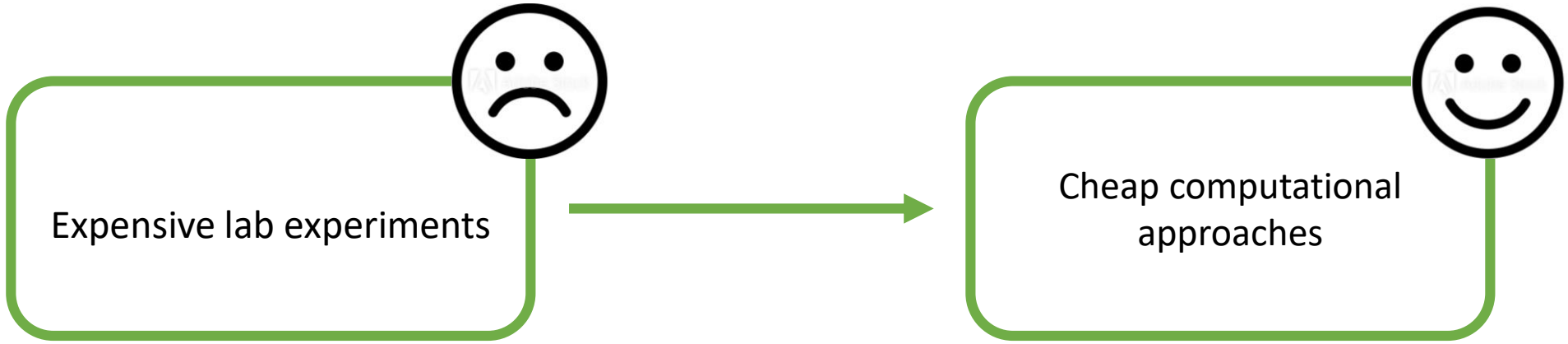
Substrate Promiscuity is important for biological engineering

- Find right enzyme to:
  - Degrade a molecule
  - Produce a molecule

→ Design synthetic metabolic pathways

→ Production of compounds for pharmaceuticals

Problem statement



Challenge: develop computational approaches to characterize promiscuous enzyme activity

Know **if**, and possibly **why** an enzyme will act on a molecule

Previous work

1. Prediction techniques for specific enzymes that are known to be broad (e.g. CYP450 enzymes)

   Yousofshahi, M., et al., *PROXIMAL: a method for Prediction of Xenobiotic Metabolism.* BMC systems biology, 2015. **9**(1): p. 94.

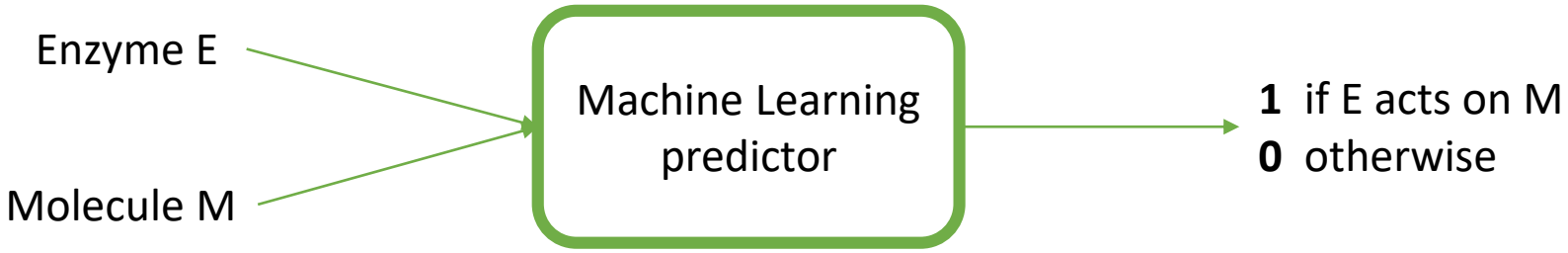
2. Prediction techniques based on similarity

   Pertusi, D.A., et al., *Efficient searching and annotation of metabolic networks using chemical similarity.* Bioinformatics, 2015. **31**(7): p. 1016-24.

3. Prediction using machine learning but limited to 4 enzymes

   Pertusi, D.A., et al., *Predicting novel substrates for enzymes with minimal experimental effort with active learning.* Metab Eng, 2017. **44**: p. 171-181.

Our data-driven approach



Binary Classification for each enzyme

Each molecule is either:

- Positive (label **1**) → (enzyme acts on it)
- Negative (label **0**) → (enzyme does not act on it)

We effectively built one predictor per enzyme

Data

- BRENDA catalogues known natural and promiscuous reactions present in the literature for over 6000 enzymes

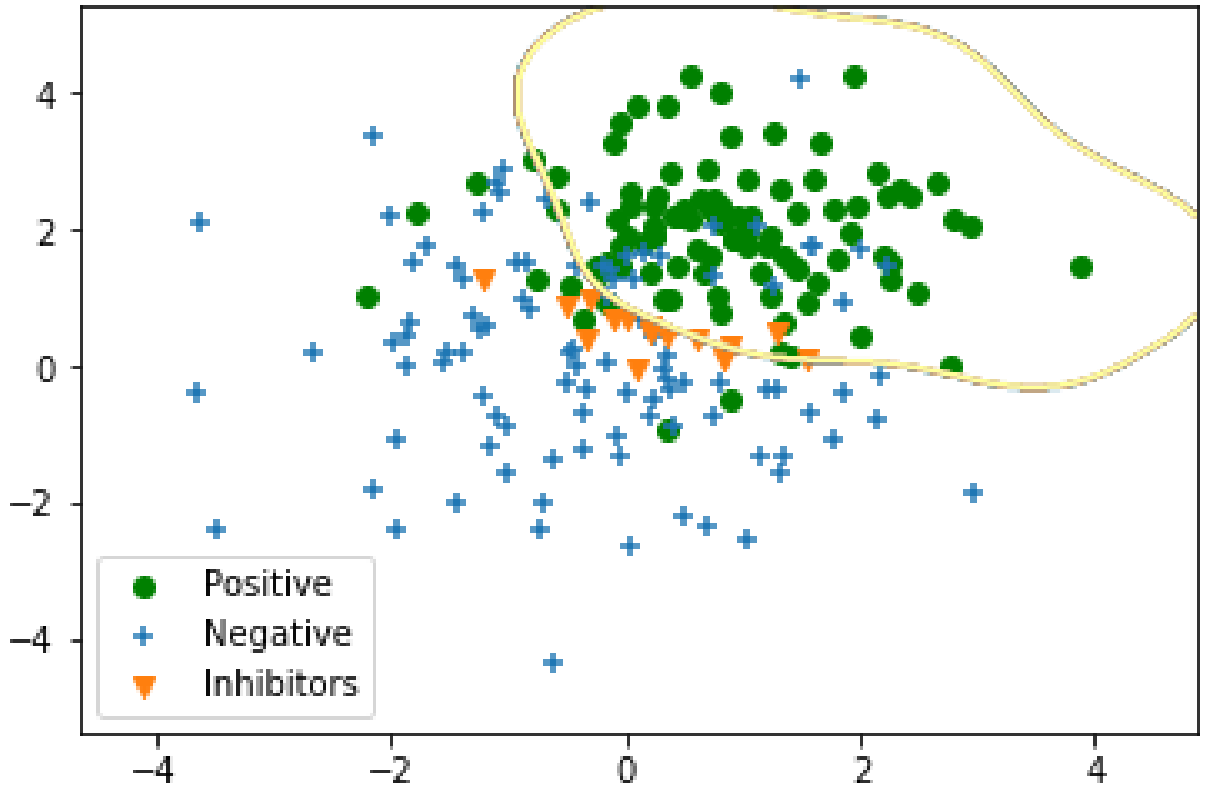
- First summer: preprocessing to get usable representations of molecules
  - Minimize loss of data

- After preprocessing: data for 1007 enzymes (1007 predictors)
  - Recall: previous work only focuses on much smaller sets of enzymes

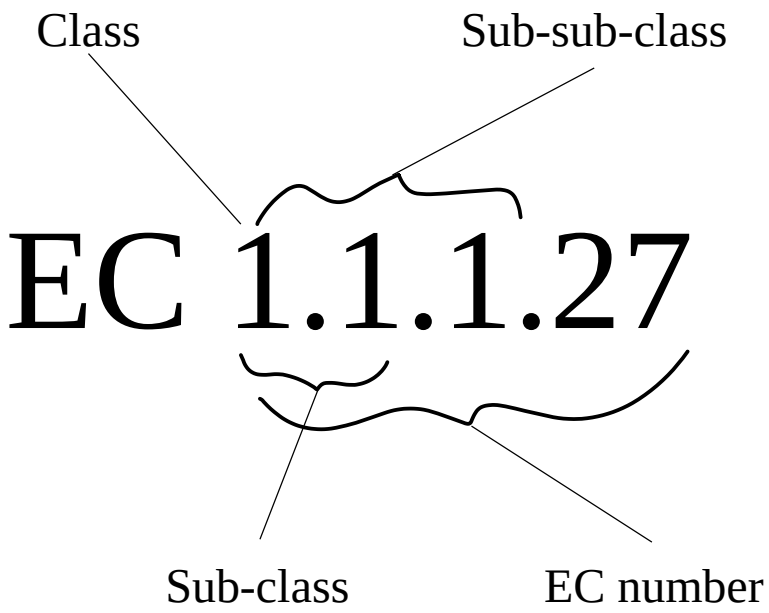
Problems with the data

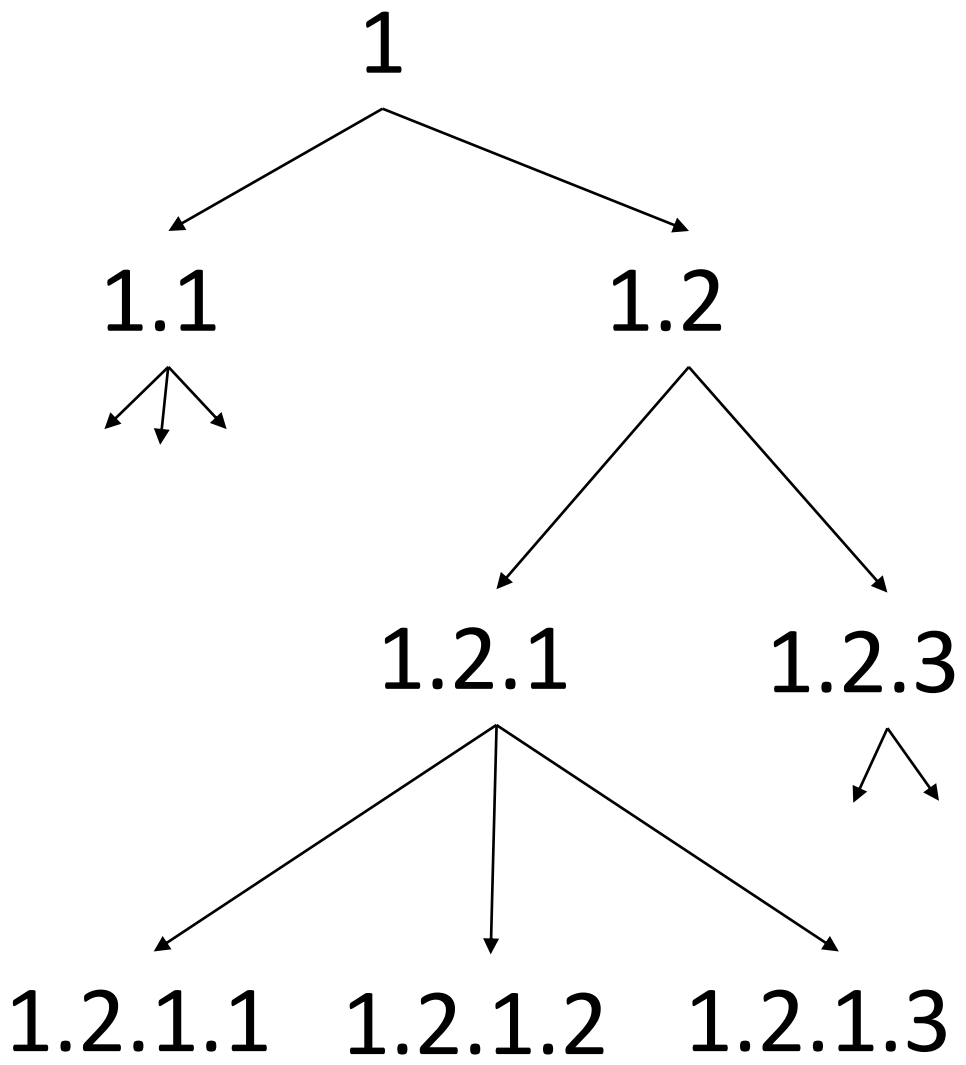
1. Many enzymes have limited data
2. No comprehensive database with negative example molecules
    - Unlabeled molecules: label is not known
    - Treat unlabeled molecules as negative
        - More negative than positive molecules in nature

1) Inhibitors as hard negative examples



2) Share information across EC hierarchy





3) Similarity-based confidence on unlabeled molecules

Unlabeled molecule $\boldsymbol{u}$
Positive set $\boldsymbol{P}$

$$confidence_u = 1 - maxSimilarity(u, P)$$

Evaluation

- We kept some data aside for testing
  - Molecules for which we know the true label
- We evaluated our model by making it predict the label of known test molecules
- Compared against predictions made by a **Baseline** model

Results

1. All our three contributions improve performance
   - Hierarchical structure delivers best improvement
2. Best model:
   - Has all three techniques
   - On average, we predict the correct answer about **92%** of the times

Conclusion

- We can now correctly predict enzyme activity for 1007 enzyme on any query molecule about **92%** of the times
  - Useful tool → will be up on a website
  - Some enzymes easier to characterize than others → **future work**
- Data-driven machine-learning approach
  - Easily scales with more data